Linguistic Resources for the Meeting Domain

Meghan Lammie Glenn, Stephanie M. Strassel

Linguistic Data Consortium
3600 Market Street, Suite 810
Philadelphia, PA 19104
{mlglenn, strassel}@ldc.upenn.edu

Abstract. This paper describes efforts by the University of Pennsylvania's Linguistic Data Consortium to create and distribute shared linguistic resources – including data, annotations, and tools – to support the Spring 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation. In addition to making available large volumes of training data to research participants, LDC produced reference transcripts for the NIST Phase II Training Corpus and RT-09 conference room evaluation set, which represent a variety of subjects, scenarios and recording conditions. For the five-hour NIST Phase II Training Corpus, LDC manually created quick transcripts which include turn segmentation and minimal markup. The three-hour evaluation corpus required the creation of careful verbatim reference transcripts including manual segmentation and rich markup. We describe the process of creating transcripts for the RT-09 evaluation, quality control, real-time transcription rates, and XTrans, LDC's next generation transcription toolkit. Finally, we present plans for further improvements to infrastructure and data collection.

Keywords: linguistic resources, transcription, annotation tools, meeting recording, conference room, XTrans

1 Introduction

Linguistic Data Consortium was established in 1992 at the University of Pennsylvania to support language-related education, research and technology development by creating and sharing linguistic resources, including data, tools and standards. Human language technology development in particular requires large volumes of annotated data for building language models, training systems and evaluating system performance against a human-generated gold standard. LDC has directly supported the National Institute of Standards and Technology's (NIST) Rich Transcription evaluation series by providing training and evaluation data and related infrastructure. For the spring 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation, LDC provided large quantities of training data from a variety of domains to program participants. In addition, LDC created five hours of new quick transcripts for the

NIST Phase II Training conference room corpus and three hours of careful reference transcripts of evaluation data to support automatic speech-to-text transcription, diarization, and speaker segmentation and localization in the meeting domain. The RT-09 conference room sets were created by using XTrans, the specialized speech annotation tool that LDC developed to respond to unique challenges presented by transcription. XTrans supports rapid, high-quality creation of rich transcripts, in the meeting domain and in a wide variety of other genres. It also provides built-in quality control mechanisms that facilitate consistency and improve real-time transcription rates.

2 Data

2.1 Training Data

To enhance availability of high-quality training data for RT-09, LDC coordinated with NIST to distribute eight corpora that are part of the LDC catalog for use as training data by evaluation participants. The data included five corpora in the meeting domain and two large corpora of transcribed conversational telephone speech (CTS) as well as one corpus of transcribed broadcast news (BN). All data was shipped directly to registered evaluation participants upon request, after sites had signed a user agreement specifying research use of the data. The distributed training data is summarized in the table below.

Title	Speech	Transcripts	Volume	Domain
Fisher English Part 1	LDC2004S13	LDC2004T19	750+ hours	CTS
Fisher English Part 2	LDC2005S13	LDC2005T19	750+ hours	CTS
ICSI Meeting Corpus	LDC2004S02	LDC2004T04	72 hours	Meeting
ISL Meeting Corpus	LDC2004S05	LDC2004T10	10 hours	Meeting
NIST Meeting Pilot Corpus	LDC2004S09	LDC2004T13	13 hours	Meeting
RT-04S Dev-Eval Meeting Room Data	LDC2005S09	LDC2005S09	14.5 hours	Meeting
RT-06 Spring Meeting Speech Evaluation Data		LDC2006E16	3 hours	Meeting
TDT4 Multilingual Broadcast News Corpus	LDC2005S11	LDC2005T16	300+ hours	BN

Table 1. RT-09 Training Data Distributed by LDC

2.2 NIST Phase II Training Data

LDC transcribed five hours of meeting recordings for the NIST Phase II Training Corpus, using the quick transcription (QTR) methodology. [1] The corpus is comprised of four files, ranging from 33-106 minutes in duration. There are three to

eight speakers per session, including native and non-native speakers. The topic content is primarily business-oriented: product presentations and demonstrations and journal article reviews.

2.3 Evaluation Data

In addition to making the training data available, LDC developed a portion of the benchmark test data for this year's evaluation. The RT-09 three-hour conference room evaluation corpus includes seven excerpts contributed by three organizations or consortia. The sessions contain four to eleven participants and are between 19 and 30 minutes long. In all cases individual head-mounted microphone (IHM) recordings were available and were used for the bulk of transcription. Senior annotators use merged head mounted microphones or distant microphones for quality control. The meetings represent a variety of subjects, scenarios and recording conditions, and content.

3 Transcription

3.1 Quick Transcription (QTR)

The goal of quick transcription (QTR) is to "get the words right" as quickly as possible; to that end, the QTR methodology eliminates most feature markup, permitting transcribers to complete a verbatim transcript in a single pass over each channel, during which transcribers segment and transcribe the audio signal into speaker turns and utterances. [1]

The QTR approach was adopted on a limited scale for English conversational telephone speech data within the DARPA EARS program [2], with real-time transcription rates of seven to ten times real-time. Team leaders monitor progress and speed to ensure that transcripts are produced within the targeted timeframe. The resulting quick transcription quality is naturally lower than that produced by the careful transcription methodology, since accelerating the process inevitably results in missed or mis-transcribed speech; this is particularly true for difficult sections of the transcript, such as disfluent or overlapping speech sections. However, the advantage of this approach for producing training data is undeniable. Annotators work ten times faster on average using this approach than within the careful transcription methodology.

3.1.1 Quality Control

Quality assurance efforts are minimized for QTR, since the goal of this approach is to produce a transcript in as little time as possible. However, the meetings in this dataset were reviewed in a brief second pass involving a spell check and a file format check.

Transcripts were reviewed again briefly by a team leader for accuracy and completeness.

3.2 Careful Transcription (CTR)

For purposes of evaluating transcription technology, system output must be compared with high-quality manually-created verbatim transcripts. LDC has already defined a careful transcription (CTR) methodology to ensure a consistent approach to the creation of benchmark data. [3] The goal of CTR is to create a reference transcript that is as good as a human can make it, capturing even subtle details of the audio signal and providing close time-alignment with the corresponding transcript. CTR involves multiple passes over the data and rigorous quality control. Some version of LDC's current CTR specification has been used to produce test data for several speech technology evaluations in the broadcast news and conversational telephone speech domains in English, Mandarin, Modern Standard and Levantine Arabic as well as other languages over the past decade. In 2004 the CTR methodology was extended to the meeting domain to support the RT-04 meeting speech evaluation.

Working with a single speaker at a time using individual head-mounted microphone (IHM) recordings, annotators first divide the audio signal into virtual segments containing speaker utterances and noise while simultaneously labeling each speaker with a unique speaker ID. At minimum, annotators divide the audio into individual speaker turns. Turns that are longer than 10 seconds are segmented into smaller units. Speaker turns can be difficult to define in general and are particularly challenging in the meeting domain due to the frequency of overlapping speech and the prevalence of side conversations or asides that occur simultaneously with the main thread of speech. Transcribers are therefore generally instructed to place segment boundaries at natural breakpoints like breath groups and pauses, typically resulting in segments of three to eight seconds in duration.

When placing segment boundaries, transcribers listen to the entire audio file and visually inspect the waveform display, capturing every region of speech as well as isolating vocalized speaker noises such as coughs, sneezes, and laughter. Transcribers leave several milliseconds of silence padding around each segment boundary so as not to clip off the onset of voiceless consonants or the ends of fricatives.

After accurate segment boundaries are in place, transcribers create a verbatim transcript by listening to each segment in turn. Because segments are typically around five seconds long, it is usually possible to create a verbatim transcript by listening to each segment once, though regions containing speaker disfluencies or other phenomena may warrant several reviews. While no time limit is imposed for CTR, annotators are instructed to use the "uncertain transcription" convention if they have reviewed a segment three or more times and are not confident in how to transcribe the utterance. A second pass by a different transcriber checks the accuracy of the segment boundaries and transcript itself, revisits sections marked as “uncertain,” validates speaker identity, adds information about background noise conditions, and inserts special markup for mispronounced words, proper names, acronyms, partial words, disfluencies and the like. A third pass over the transcript conducted by the team leader

ensures accuracy and completeness, leveraging the context of the full meeting to verify specific vocabulary, acronyms and proper nouns as required.

Transcription ends with multiple automatic and manual scans over the data to identify regions of missed speech, fix common errors, expand contractions, and check the file format. These steps are described in more detail in the following section.

3.2 Quality Control

To enhance the accuracy of meeting transcription, annotators work with the separate IHM recordings of individual speakers and the merged recording of the all IHM recordings of the meeting participants. Segmentation and first-pass transcription are produced primarily from the individual IHM recordings in the manner described above.

Meetings may contain highly specialized terminology and names that may be difficult for transcribers to interpret. To resolve instances of uncertainty and inconsistency, senior transcribers conduct an additional quality control pass, using a distant or table-top microphone recording or the merged IHM recording to obtain a comprehensive overview of a discussion. During this additional pass, senior annotators also check for common errors and standardize the spelling of proper nouns and the representation of acronyms in the transcript and across transcripts, where applicable.

The last stages of quality control involve multiple quality assurance scans such as listening to all untranscribed regions of individual recordings to identify any areas of missed speech or chopped segments. Finally, annotators check spelling and file format, and expand contractions.

4 Unique Challenges of Meeting Data

The meeting domain presents a number of unique challenges to the production of highly accurate verbatim transcripts, which motivates the multi-pass strategy described above. One challenge is the prevalence of overlapping speech, which in this domain is fairly frequent, accounting for approximately 25% of the speech on average.[1] Even when the transcriber has muted all but one speaker's IHM recording, accurately transcribing speech in overlapping regions can be difficult because other speakers are typically still audible. During all stages of transcription, transcribers and team leaders devote careful attention to overlapping speech regions.

Meeting content may also present a challenge to transcribers. Much of the conference room data is collected during project discussion groups or technical meetings, and frequently involves highly-specific terminology or acronyms that motivate extra care and research to transcribe accurately. For example, the term "WIIFM" was transcribed as "with them" until a transcriber, prompted by the context of the utterance, confirmed its use as an acronym for "What's In It For Me." Terms like this may require one or two independent transcription passes to understand.

[1] This is based on the RT-09 test set, where the amount of overlap ranged from 4.85%-43.04%.

Another challenge fundamental to creating a high-quality meeting data transcripts is the added volume of speech, resulting from not one or two but a half a dozen or more speakers. A typical thirty-minute telephone conversation will require twenty hours or more to transcribe carefully (30 minutes, two speakers, 20 times real-time per channel). A meeting of the same duration with six participants may require more than 60 hours to produce a transcript of the same quality.

The nature of meeting speech transcription requires frequent jumping back and forth from a single speaker to a multi-speaker view of the data, which presents a challenge not only for the transcribers, but for the transcription tools they use. Many current transcription tools are not optimized for or do not permit this approach. For the most part existing transcription tools cannot incorporate output of automatic processes, and they lack correction and adjudication modes. Moreover, user interfaces are not optimized for the tasks described above.

5 Infrastructure

LDC has been using a next-generation speech annotation toolkit, XTrans, to directly support a full range of speech annotation tasks including quick and careful transcription of meetings since late 2005.

The current version of XTrans runs on FreeBSD, Linux and Windows platforms. Most of the XTrans components are written in Python with some components written in C++, such as the QWave waveform display module, based on the Qt GUI toolkit. It contains customized modules for quick and careful transcription and structural spoken metadata annotation. The tool supports bi-directional text input, a critical component for languages such as Arabic. XTrans is being used for full-fledged transcription and a variety of speech annotation tasks in Arabic, Mandarin Chinese, and English at LDC.

XTrans contains user-configurable key bindings for common tasks. All commands can be issued from keyboard or mouse, depending on user preference. This user-friendly tool includes specialized quality control features; for instance, speakerID verification, which help transcribers identify misapplied speaker labels by incorporating commands to listen to random segments – or all segments – of one speaker. In addition, XTrans includes silence checking to identify speech within untranscribed regions. XTrans enables easy handling of overlapping speech in single-channel audio by implementing a Virtual Speaker Channel (VSC) for each speaker, not each audio channel. This is particularly useful for the final quality control stages of careful transcription, when senior annotators review the transcript by listening to the merged head-mounted microphone recording.

To support meeting domain transcription, XTrans permits an arbitrary number of audio channels to be loaded at once. For RT-09, transcribers created the transcripts by listening to the IHM channels for each meeting recording session. They had access to distant microphone recordings when desired, and easily toggle between the multi- and single-speaker views, turning individual channels on and off as required to customize their interaction with the data. The waveform markup display makes speaker interaction obvious, showing overlapping segments and assigning a unique color to

each speaker. Overlapping segments among speakers are shown by the overlapping color coded boxes in the waveform display. The audio and transcript are linked, so that clicking on a segment in the transcript will highlight the corresponding region in the appropriate audio file.

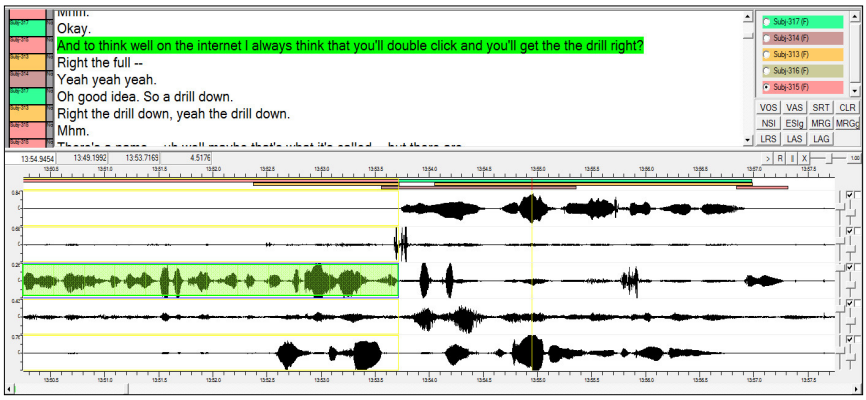

Figure 1. Meeting recording session with five speakers, as shown in XTrans.

When the transcriber desires a multi-speaker view of the meeting session, the transcript for all meeting participants is shown in the text edit window. The transcriber can activate the sound for the all recordings by toggling the radio buttons next to the waveform for each channel. When the transcriber desires a more focused view of the meeting, he may show the segments for just one speaker, muting the audio recordings of the other meeting participants accordingly by de-selecting the audio output buttons to the other audio channels.

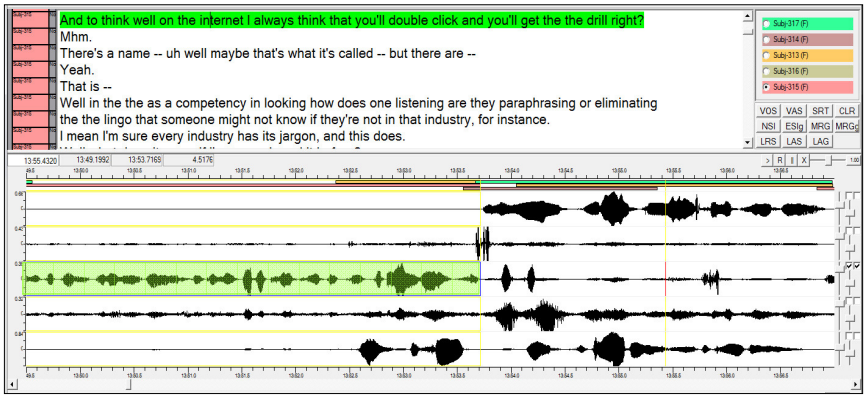

Figure 2. Meeting recording session with five speakers, focus on one speaker, as shown in XTrans.

As with LDC's current annotation tools, XTrans is fully integrated into LDC's existing annotation workflow system, AWS. AWS controls work (project, file) assignment; manages directories and permissions; calls up the annotation software and assigned file(s) for the user; and tracks annotation efficiency and progress. AWS allows for double-blind assignment of files for dual annotation, and incorporates adjudication and consistency scoring into the regular annotation pipeline. Supervisors can query information about progress and efficiency by user, language, data set, task, and so on.

6 Transcription Rates

LDC careful transcription real-time rates for the RT-05S two-hour dataset approached 65 times real-time, meaning that one hour of data required around 65 hours of labor (excluding additional QC provided by the team leader), which is around 15 times real-time per channel [6], comparable with rates for BN and slightly less than that for CTS. Using XTrans to develop the RT-06S conference room data, the team's real-time rates dropped to under 50 times real-time per file (10 times real-time per channel), remained steady through RT-07 [7], and were approximately 40 times real time per file (approximately eight times real time per channel) for the RT-09 test data.

7 Future Plans and Conclusion

LDC's planned activities include data collection in the meeting domain, Using existing facilities at LDC developed for other research programs, meeting collection is currently opportunistic, with regularly scheduled business meetings being recorded as time allows. As new funding becomes available, we also plan to develop our collections infrastructure with additional head-mounted and lavaliere microphones, an improved microphone array, better video capability and customized software for more flexible remote recording control. While the current collection platform was designed with portability in mind, we hope to make it a fully portable system that can be easily transported to locations around campus to collect not only business meetings but also lectures, training sessions and other kinds of scenarios.

LDC plans to explore inter-transcriber consistency for this domain, and develop segmentation and annotation methods that would enhance the quality or value of reference meeting transcripts.

Future plans for XTrans include incorporation of video input to assist with tasks like speaker identification and speaker turn detection. We also plan to update the text widget to allow layers of annotation over the text transcript, for richer disfluency annotation. In addition, we plan to add a "correction mode" that will allow users to check manual transcripts or verify output of automatic processes including auto-segmentation, forced alignment, SpeakerID and automatic speech recognition output. Another XTrans feature which we would like to implement is the "adjudication mode", allowing users to compare, adjudicate and analyze discrepancies across

multiple human or machine-generated transcripts. This function would directly support our goal of studying inter-transcriber consistency.

Shared resources are a critical component of human language technology development. LDC is actively engaged in ongoing efforts to provide crucial resources for improved speech technology to RT-09 program participants as well as to the larger community of language researchers, educators and technology developers. These resources are not limited to data, but also include annotations, specifications, tools and infrastructure.

Acknowledgments. We would like to thank Haejoong Lee for his development and maintenance of the XTrans speech annotation tool. We would also like to thank the LDC transcription team for their hard work in creating the transcripts for RT-09S.

8 References

1. Linguistic Data Consortium: RT-09 Meeting Quick Transcription Guidelines. (2009) http://projects.ldc.upenn.edu/Transcription/MeetingRecording/MeetingQTR-V3.0.pdf
2. Strassel, S., Cieri, C., Walker, K., Miller, D.: Shared Resources for Robust Speech-to-Text Technology, Proceedings of Eurospeech (2003).
3. Linguistic Data Consortium: RT-09 Meeting Careful Transcription Guidelines. (2009) http://projects.ldc.upenn.edu/Transcription/MeetingRecording/MeetingCTR-V2.3.pdf
4. Bird, S., Liberman, M: A formal framework for linguistic annotation. Speech Communication, (2001) 33:23-60.
5. Maeda, K., Strassel, S. M.: Annotation Tools for Large-Scale Corpus Development: Using AGTK at the Linguistic Data Consortium. Proceedings of the 4th International Conference on Language Resources and Evaluation (2004).
6. Glenn, M., Strassel, S. M.: Linguistic Resources for Meeting Speech Recognition. MLMI 2005: 390-401
7. Glenn, M., S. M. Strassel, "Shared Linguistic Resources for the Meeting Domain," in Stiefelhagen, R., R. Bowers, and J. Fiscus, (Eds.), in Lecture Notes in Computer Science, vol. 4625, Multimodal Technologies for Perception of Humans, Heidelberg: Springer, 2008, pp. 401-413.